

JP2001067187

Biblio Page 1





STORAGE SUB-SYSTEM AND ITS CONTROL METHOD

Patent Number:

JP2001067187

Publication date:

2001-03-16

Inventor(s):

ARAKAWA TAKASHI; MOGI KAZUHIKO; YAMAKAMI KENJI; ARAI

HIROHARU

Applicant(s)::

HITACHI LTD

Requested Patent:

JP2001067187 (JP01067187)

Application

Number:

JP19990242713 19990830

Priority Number(s):

IPC Classification:

G06F3/06; G06F12/00

EC Classification:

Equivalents:

Abstract

PROBLEM TO BE SOLVED: To simplify a work for optimizing arrangement by re-arrangement by the user of a disk array system or the like by changing the correspondence of a logical storage area from a physical storage area into the second physical storage area and executing re-arrangement.

SOLUTION: A control part 300 automatically executes re-arrangement execution processing at the set time and date. That is, the part 300 copies contents stored in a re-arrangement source physical area in a re-arrangement destination physical area based on re-arrangement information 408. Moreover, at the point of time when the copying is completed and the whole contents of the re-arrangement source physical area are reflected in the re-arrangement destination physical area, the control part 300 changes a physical area corresponding to a logical area for executing re-arrangement in logical/physical correspondence information 400 from the re-arrangement source physical area into the re-arrangement destination physical area. Besides, the control part 300 uses the re-arrangement destination physical area on a non-usage physical area 1470, changes the re-arrangement source physical area into the non-usage one and, moreover, updates the time and date of re-arrangement execution time information 406 into the one for a next time by referring to time and date updating information on re-arrangement execution time information 406.

Data supplied from the esp@cenet database - I2

50 超えないように決定する手段とを有することを特徴とす

開特許公報(A) (22) (18) 日本日本日本日(1b)

特開2001-67187 (11)特許出國公開番号

(P2001-67187A)

(43)公開日 平成13年3月16日(2001.3.16)

(事業)	3014 58065		_
PI	G06F 3/06		12/00
中温高温	301	540	501
51)Int.Cl.'	C06F 3/06		12/00

(金24 里) 育型開次 未開次 競次項の数10 〇1

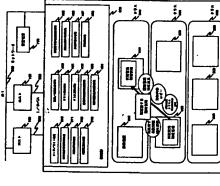
(21)出題等中	特閣平11-242713	(元) 出題人	(71)出版人 000005108	
(22) 出口日	平成11年8月30日(1999.8.30)		株式会社日立製作所 第古表千代田区第四聯盟会四丁日 s 争体	
		(72) 完明者	加二 表文	
			神疾川県川崎市麻生区王禅寺1089番地 株	
			式会社日立製作所システム開発研究所内	
		(72) 発明者	茂木 和蘇	
			神奈川県川崎市麻生区王禅寺1099番地 徐	
			式会社自立製作所システム開発研究所内	
		(74)代理人		
			非理士 作田 康夫	

(54) 【発明の名称】 ストレージサブシステム及びその観響方法

(21) [財政]

[歌昭] ストレージサブシステムのユーザまたは保守員 が記憶領域の物理的再配置による配置最適化を行うため の作策を節便にするストレージサブシステムおよび制御 方法を提供する。

【解決手段】ストレージサブシステム200は、記憶装 600として管理し、クラス属性に基づき好適な再配置 **買500を、それぞれ属性を有する複数の組(クラス)** 先のクラスを決定する。



|特許額水の範囲|

|静水項1||複数の配億装置と、前記記憶装置の使用状 祝僧報を取得する手段と、前記計算機がリードライト対 領域との対応づけを行う手段とを有し、1台以上の計算 機に接続するストレージサブシステムの制御方法であっ て、前記記憶装置は複数の組(クラス)に分類され、前 システムは、前記使用状況情報および前記クラス属性に 基づき前記論理記憶領域に好適な再配置先のクラスを決 の物理記憶領域から前記第二の物理記憶領域へ変更して 象とする論理記憶倒域と前記記憶装置の第一の物理記憶 定し、前記論理記憶領域の再配置先として利用可能な第 の物理記憶領域の内容を前記第二の前記物理記憶領域に 二の物理記憶領域を前記クラス内から選択し、前記第一 **リアーナるとかもに種邸的勧励機の対抗 かけかぎ的終一** 再配置を行うことを特徴とする前記ストレージサブシス 記クラスは設定された属性を有し、前記ストレージサン

使用状況情報を蓄積し、散定された期間の前記使用状況 の制御方法であって、ストレージサブシステムは、前記 情報に基づき、論理記憶領域の再配價先を決定し、設定 された時間に再配置を行うことを特徴とするストレージ 【請求項2】請求項1に記載のストレージサブシステム サブシステムの制御方法。

システムの制御方法であって、ストレージサブシステム 超えている記憶装置から再配置する論理記憶領域を選択 は、使用状況情報として、記憶装置の単位時間当たりの 使用時間 (使用率) を用い、各クラスは、属性として設 定されたクラス間の性能順位と使用率上限値を有し、前 し、前記論理記憶領域の再配置先のクラスを前記類位の 上位のクラスから、各クラスの使用率上限値を超えない 【請求項3】請求項1または2に記載のストレージサブ 記ストレージサブシステムは、クラスの使用単上限値を ように決定することを特徴とするストレージサブシステ

記ストレージサブシステムは、クラスの使用率上限値を 【請求項4】 請求項1または2に記載のストレージサン システムの制御方法であって、ストレージサブシステム は、使用状況情報として、記憶装置の単位時間当たりの 使用時間(使用率)を用い、各クラスは、属性として設 定されたクラス間の性能顕位と使用率上限値を有し、前 組えている記憶装置から再配置する論理記憶領域を選択 し、前記論理記憶領域の再配置先として利用可能な物理 記憶領域を同一クラス内の記憶装置から、前記クラスの 使用率上限値を超えないように決定することを特徴とす るストレージサブシステムの制御方法。

使用時間(使用率)を用い、各クラスは属性として設定 システムの制御方法であって、ストレージャブシステム は、使用状況情報として、記憶装置の単位時間当たりの された対象アクセス種別と使用率上限値を有し、前記ス 【請水項5】請水項1または2に記載のストレージサン

前記論理記憶領域に対するアクセス種別の分析結果に基 づいて前記論理記憶領域の再配置先のクラスを前記対象 アクセス種別のクラスから、各クラスの使用車上限値を 組えないように決定することを特徴とするストレージサ トレージサブシステムは、クラスの使用率上限値を超え ている記憶装置から再配置する論理記憶領域を選択し、 インステムの制御方法。

梅服2001-67187

【請求項6】1台以上の計算機に接続し、複数の記憶装 匠と、前記記憶装置の使用状況情報を取得する手段と、 10 前記計算機がリードライト対象とする陥埋記憶領域と前 記記憶装置の第一の物理記憶倒域との対応づけを行う手 段とを有するストレージサブシステムであって、前記核 ラス)として管理する手段と、前記使用状況情報および 前記クラス異性に基づき前記論理記憶領域に好適な再配 容を前記第二の前記物理記憶領域にコピーするとともに 前記第二の物理記憶領域へ変更して再配配を行う手段と 数のディスク装置をそれぞれ属性を有する複数の組(ク **置先のクラスを決定する手段と、前記論理記憶領域の再** 配配先として利用可能な第二の物理記憶倒域を前記クラ **福理記憶領域の対応

カナを前記第一の物理記憶図域から** ス内から選択する手段と、前記第一の物理記憶倒域の内 2

【請求項1】請求項6に記載のストレージサブシステム であって、ストレージサブシステムは、帕記使用状改情 報を蓄積し、歓定された期間の前記使用状況情報に基づ と、数定された時間に再配数を行う手段とを有すること き、論理記憶領域の再配配先を自動的に決定する手段 を有することを特徴とするストレージサブシステム。 52

を作数とするストレージサンシステム。

システムであって、ストレージサブシステムは、使用状 **祝情報として記憶装費の単位時間当たりの使用時間(使** 用率)を用いる手段を有し、前記ストレージサブシステ ムは、各クラスに属性として設定されている使用事上限 値を組えている記憶装置から再配置する論理記憶領域を 選択する手段と、前記論理記憶関域の再配置先のクラス 頃位から、各クラスの使用率上限値を超えないように決 【酢水項8】 肺水項6または7 に記載のストレージサン を各クラスに属性として設定されているクラス間の性能 定する手段とを有することを特徴とするストレージサブ ೫ 33

システムであって、ストレージサブシステムは、使用状 【酵水項9】 精水項6または7に記載のストレージサブ 祝情報として、記憶装置の単位時間当たりの使用時間 システム。 \$

ステムは、属性として設定されたクラスの使用率上限値 を超えている記憶装置から再配置する論理記憶節域を選 択する手段と、前配論理記憶領域に対するアクセス種別 を分析する手段と、対象アクセス種別を属性として設定 されたクラスから、前記論理記憶領域の再配置先のクラ スを前配分析結果に基づいて各クラスの使用申上限値を (使用率) を用いる手段を有し、前記ストレージサブシ

るストレージサブシステム。

[請求項10] 請水項6、7、8、または9に記載のス り、前記ディスク装匠の使用率を使用状没情報として用 トレージサブシステムであって、ストレージサブシステ ムは、複数のディスク装置を有するディスクアレイであ いろ手段を有することを特徴とするストレージサブシス

【発明の評価な説明】

[0001]

[発明の属する技術分野] 本発明は、複数の記憶装置を 作するストレージサブシステム、およびその制御方法に

[0002]

を実現する二次記憶システムの1つにディスクアレイシ ステムである。ディスクアレイシステムに関する論文と 【従来の技格】コンピュータシステムにおいて、高性能 LTH. D. A. Patterson, G. Gibs ステムがある。ディスクアレイシステムは、複数のディ スク技限をアレイ状に配置し、前配各ディスク装置に分 別格納されるデータのリードノライトを、前記各ディス ク装匠を並列に動作させることによって、高速に行うシ for Redundant Arrays of I 109-116, June 1988) 1885, 30 スクアレイシステムをレベル0と呼ぶこともある。上記 (in Proc. ACM SIGMOD, pp. 鉛文では、冗長性を付加したディスクアレイシステムに **対し、その構成に応じたフペケ1かむフペル5の種別を 序えている。これらの極別に加えて、冗長性無しのディ** のキレベルは冗長性などにより実現するためのコストや 性能特性などが異なるため、ディスクアレイシステムを **構築するにあたって、複数のレベルのアレイ (ディスク** 装尻の机)を混在させることも多い。ここでは、この組 and R. H. Kats, "A Case ncxpensive Disks (RAID)" のことをパリティグループと呼ぶ。

[0003] ディスク装頂は、性能や容品などによりコ ストが異なり、ディスクアレイシステムを構築するにあ やはり性能や容量の異なる複数種のディスク装置を用い たって最適なコストパフォーマンスを実現するために、

他知域とディスク装置の記憶関域を示す物理記憶質域の ムに接続するホストコンピュータがアクセスする論理記 対応づけ(アドレス変換)を行う。特別平9-2745 4.4.5公報には、ホストコンピュータからの論理記憶質 域に対する1/0アクセスについての情報を取得する手 段と、論理記憶領域の物理記憶領域への対応づけを変更 して物理的再配置を行う手段により、格納されたゲータ {0004} ディスクアレイシステムに格納されるデー め、ディスクアレイシステムは、ディスクアレイシステ タを上記のようにディスク装置に分散して配置するた

の最適配置を実現するディスクアレイシステムが開示さ

[0005]

4 号公報に示されるような従来の技術における配置最適 [発明が解決しようとする課題] 特開平9-27454 化の実行方法については以下の課題がある。 9

ステムの構成や個々のディスク装置の特性や性能などの 情報を確認して前記選択を行わなければならず、ユーザ 【0006】 再配置する論理記憶領域の選択および再配 **歴先の物理記憶領域の選択にあたり、ディスクアレイシ** ステムのユーザまたは保守員が、前配ディスクアレイシ または保守員による作業が煩雑となっていた。 2

【0007】また、ディスクアレイシステムが選択を自 動的に行う場合においても、ユーザまたは保守員が前記 個々のディスク装置の情報を確認して選択基準値を規定 しなければならず、やはりユーザまたは保守員による作 薬が煩雑となっていた。特に、上記のように異種のレベ ルや異種のディスク装匠の混在するディスクアレイシス テムについては情報管理の煩雑さが増大する。 2

【0008】また、ディスクアレイシステムが選択のた めに行う 1 /Oアクセス情報の参照は、ホストコンピュ 一タおよびディスクアレイシステムを含むシステムで行 われる処理のスケジュールの特性を考慮していなかっ ឧ

[0015]

た。一般にコンピュータシステムで行われる処理と処理 ルに則って行われており、また処理および1/0の傾向 は日毎、月毎、年毎などの周期性を示す場合も多く、一 役にユーザは特定期間の処理および1/0に関心がある に伴う 1 /0は、ユーザによって作成されたスケジュー と考えられる。

理的再配置による性能チューニング方法は、ディスク装 [0009]また上記従来技術において、再配置による 性能チューニング方法については以下の親題がある。物 登、すなわち、物理記憶領域の使用状況に変更を加える ものであるが、従来の技術においては、ホストコンピュ ータからの倫理記憶卸版に対する1/0アクセスについ ての情報を参照するため、再配置する論理記憶領域の選 択および再配置先の物理記憶領域の遊択にあたり、正し い選択が行えない可能性があった。 8 33

シャルアクセスとランダムアクセスが顕著に、同一のデ セスを異なるディスク装置に分離するために、再配置先 らの処理要件として、データ長の小さいランダムアクセ クセスが存在する場合、ランダムアクセスの応答時間は 【0010】また、ホストコンピュータからのシーケン イスク装置に含まれる別々の物理記憶領域に対して行わ のディスク装置を任意に特定して自動的再配置を行わせ ることはできなかった。一般に、ホストコンピュータか れる場合でも、シーケンシャルアクセスとランダムアク 同一ディスク装置にデータ長の大きいシーケンシャルア スには短時間での応答(高応答性能)が求められるが、 シーケンシャルアクセスの処理に阻害されて長くなり、 **\$** S 45

【0011】本発明の第一の目的は、ディスクアレイシ **応答性能は悪化してしまう。**

ステムのユーザまたは保守員が再配置による配置最適化 【0012】本発明の第二の目的は、ホストコンピュー タおよびディスクアレイシステムを含むシステムでの処 理のスケジュールを考慮した再配置による配置最適化を を行うための作業を簡便にすることにある。 可能にすることにある。

憶領域の選択および再配置先の物理記憶領域の選択にあ **基乙く強択を行う、ディスクアワイシステムの制御方法** 【0013】本発明の第三の目的は、再配置する論理記 **ステムにおける同一ディスク装置での顕著なシーケンシ** ヤルアクセスとランダムアクセスの混在に対し、再配置 たり、実際の記憶装置であるディスク装置の使用状況に 【0014】本発明の第四の目的は、ディスクアレイシ 先のディスク装置を任意に特定して再配置によりシーケ ンシャルアクセスおよびランダムアクセスを異なるディ スク装置に自動的に分離することができるようにするこ およびディスクアレイシステムを提供することにある。 とにある。

イスクアレイシステムは、配下の複数のディスク装置の 使用状況情報を取得する手段と、ホストコンピュータが る複数の組(クラス)として管理する手段と、使用状況 配置先のクラスを決定する手段と、論理記憶領域の再配 【課題を解決するための手段】上記の第一目的を実現す るために、1台以上のホストコンピュータに接続するデ リード/ライト対象とする論理記憶領域とディスク装置 し、さらに、複数のディスク装置をそれぞれ属性を有す 情報およびクラス風性に基づき論理記憶領域に好適な再 置先として利用可能な第二の物理記憶領域をクラス内か ら選択する手段と、第一の物理記憶領域の内容を前記第 二の前記物理記憶領域にコピーナるとともに論理記憶質 核の対応ムけを第一の物理記憶質域から第二の物理記憶 の第一の物理記憶簡減との対応のけを行う手段とを有 領域へ変更して再配置を行う手段を備える。

ディスクアレイシステムは、使用状況情報を蓄積し、設 定された期間の使用状況情報に基づき、論理記憶領域の 再配置先を決定する手段と、設定された時間に再配置を 【0016】また、上配第二の目的を実現するために、 行う手段を備えることができる。

スク装置の単位時間当たりの使用時間(使用率)を用い ディスクアレイシステムは、使用状況情報として、ディ 【0017】また、上記第三の目的を実現するために、 る手段を備える。

クセス種別) と使用率上限値を用いて、クラスの使用率 上限値を超えている記憶装置から再配置する論理記憶領 ディスクアレイシステムは、各クラスに属性として設定 された対象アクセス種別(シーケンシャル/ランダムア [0018]また、上配第四の目的を実現するために、

城を避択し、韓理記憶領域に対するアクセス預別の分析 結果に基づいて論理記憶領域の再配置先のクラスを好適 なアクセス程別のクラスから、各クラスの使用申上限値 を超えないように決定する手段を仰える。

【発明の実施の形態】以下、本発明の実施の形態を図1 ~因27を用いて説明する。

[0019]

છ

クラス600に基づく再配置の判断と、再配置判断およ [0020] <第一の実施の形態>本実施の形態では、 び実行のスケジューリングについて説明する。

【0021】図1は、本発明の第1の実施の形態におけ る計算機システムの構成図である。

ホスト100、ストレージサグシステム200、飽害為 【0022】本実施の形態における計算機システムは、 末700を有してなる。 [0023] ホスト100は、ストレージサブシステム サブシステム200に対しリードやライトの1/0を行 う。1/0の際、ホスト100は、ストレージサンシス 200に1/0パス800を介して接続し、メトレージ

20 デム200の記憶領域について結理領域を指定する。1 /Oパオ800の例としては、ESCON、SCSI、 ファイバチャネルなどがある。

[0024] ストレージサブシステム200は、勉御的 00は、リード/ライト処理310、使用状況情報取得 処理311、再配配判断処理312、及び再配置実行処 理313を行う。また、ストレージサブシステム200 300および複数の配憶装置500を有する。制御部3 は、論理/物理対応情報400、クラス構成情報40

1、クラス属性情報402、路理衛城使用状況情報40 悄報405、再配置実行時刻情報406、未使用領域債 3、物理饭饭使用状况格银404、再配配判断对象期間 報407、及び再配置情報408を保持する。

[0025] ホスト100、制御節300、および制御 端末100は、ネットワーク900で核税される。ネッ トワーク900の例としては、FDDI、ファイバチャ

[0026] ホスト100、慰御師300、および超御 端末100には、各々での処理を行うためのメモリ、C ネルなどがある。

それぞれ存在するが、本実施の形態の説明においては重 PUなど、計算機において一般に用いられる構成要素も [0027] ホスト100が、ストレージサブシステム 要でないため、ここでは説明を省略する。 \$

200に対してリード/ライトを行う場合のリード/ラ イト処理310、および使用状況情報取得処理311に ついて図2で説明する。

【0028】リードノライト処理310において、ホス 00に対しリードまたはライトを簡単的域を指定して受 水する (ステップ1000)。 要求を受領した制御邸3 ト100は、ストレージサグツステム2000短笛筒3 50 00は、4個用人物理対応情報400を用いて4個限額に

[0029] 論理/物理対応情報400の一個を図3に示す。論理アドレスはホスト100がリード/ライト処理310で用いる論理関域を示すアドレスである。物理アドレスは実際にデータが格納される記憶装置500上の関域を示すアドレスであり、記憶装置番号は配め記憶装置500上状のウェナ。記憶装置やアドレスは記憶装置500内での記憶質後を示す。記憶装置やアドレスは記憶装置500内での記憶質を示すアドレスである。

[0030] 次に、使用状況情報取得処理311において制御部300は、リード/ライト処理310においてリード/ライト処理310においてリード/ライト処理310においてリード/ライト処理310に対して後他用状況情報403と、リード/ライト処理310で使用した物理部域についての物理関域使用状況情報404世所が近になる(ステップ1030、1040)。 静理関域使用状況情報4041、例えば使用頻度、使用率、リード/ライトに関する。 4は、例えば使用頻度、使用率、リード/ライトに関する。 4は、例えば使用頻度、使用率、リード/ライトに関するが低に関する情報である。 4時間域を用状況情報403 はよび物理関域使用状況情報403 はよび物理関域使用状況情報403 はよび物理の感性の表現する。

[0031] 次に、胡抑即300が行う再配配判断処理 312について図4で説明する。

[0032] 記憶装置500は、ユーザによって、または切別状態として複数の相(クラス600) に分類されており、グラス600への分類はクラス棉成信仰401に設定されている。さらに、各クラス600は、ユーザによって、または初期条件として成性を設定されているり、属性は、クラス局性情報402に設定されている。クラス局性情報402は、許容使用状況や好路な使用状況やクラス同性情報402は、許容使用状況や分路体情報402の具体的な例は、以降の実施の形態で説明する。再配度判断外外側に、以降の実施の形態で説明する。再配度判断外外側に、以降の実施の形態で説明する。再配度判断外外側には、以降の実施の形態で説明する。再配度判断外外側には、以降の実施の形態で説明する。再配度判断外外側間に可能に対断が処理312の対象とする使用状況情報405には、ユーザによってまたは初期条件をして可配に対断を異常してある。

[0033] 再位置判断対象期間情報405の一個を図5に示す。別始日時から終了日時までの期間が対象期間となる。期間更新情報は大国の対象期間の設定条件であり、例えば毎週、毎日、X時間後などがありうる。超期 部300は、対象期間の警理関域使用状況情報403封

よび物理領域使用状改情報404全参照し (ステップ1100)、クラス属性情報402の各クラス600の許容使用状改などと比較して (ステップ1110)、物理的再配置を行うべき論理領域を選択する (ステップ11

[0034] さらに、柳御郎300は、クラス腐性情報402の許容使用状況や好適な使用状況やクラス間優先順位などを参照して(ステップ1130)、論理領域の再配置先のクラス600を選択し(ステップ114

0)、さらに、クラス600に属する記憶装置5000 中から論理関域の再配置先として未使用の物理関域を選択し(ステップ1150)、選択結果を再配置情報408に出力する(ステップ1160)。

[0035] 再配置付額408の一例を図6に示す。論理模域は、再配置する論理領域であり、再配置元物理領域は、論理領域に対応する現在の物理領域を示す記域数 図番号と記憶装置内アドレスであり、再配置先物理領域は、再配置先の理領域を示す記憶装置番号と記憶装置内アドレスである。図6に示すように再配置の立案は一20 少以上行われうる。ざらに制御部300は、再配置判断対象期間情報405の対関更新情報を参照して、再配置判断対象期間情報405の対象期間を次回分に更新する(ステップ1170)。上記の処理において制御部300は、路理/物理対応情報400を用い、また前記の未300は、路理/物理対応情報400を用い、また前記の未300は、路理/物理対応情報400を用い、また前記の未3000は、路理/物理対応情報400を用い、また前記の未3000は、路理/物理が表

[0036] 未使用額域情報 407の一例を図7に示す。配貸装置番号は個々の記憶装置 50を示す。記憶装置4万ドレスは記憶装置 500内での商域を示す下レスである。記憶装置番号および装置内アドレスは物理領域を示し、使用/未使用の項目は、物理回域の使用/未使用の区別を示す。制御部300は、通常、再配置判断処理312を対象期間以後、後述の再配置実行処理313以前に自動的に行う。

35 【0037】 女に、。例算部300が行う再配置実行処理 313について図8で説明する。

【0038】再配置実行時刻情報406にはユーザによってまたは初期条件として再配度実行処理313を行う日時と日時更新情報が設定されている。

40 [0039] 再起價支行時刻情報406の一例を図りに示す。 制御的309は、設定された日時に以下に説明する再配置実行処理313を自動的に実行する。 日時更新情報は次回の再配置実行処理313を行う日時の設定条件であり、例えば毎週、毎日、X時間後などがありう

45 る。 制御師300は、再配置情報408に基づき再配置 元物理領域に格納している内容を再配置先物理領域にコ ピーする (ステップ1200)。 さらに、コピーが完了 して再配置元物理領域の内容が全て再配置先物理領域に 反映された時点で、関御師300は、韓理/物理対応情 50 報400上の再配置を行う韓理領域に対応する物理領域

を再配置元物理領域から再配置先物理領域に変更する(ユモニナ・シェン)。

(ステップ1210)。 【0040】さらに、制御部300は、未使用物理領域

[0040] さらに、制御節300は、未使用物理領域470上の再配置先物理領域を使用とし、再配置元物理領域を使用ととのできた。 おらに制御部300は、再配置実行時刻情報406の日時更新情報を参照して、再配置実行時刻情報406の日時を予回分に更新する(ステップ1230)。

【0041】ユーザまたは保守員は、耐御部30のが上記の処理で用いている各債報を、耐御端末700からネットワーク900を介して、またはホスト100からネットワーク900または1/Oバス800を介して設定および確認すること、特に、再配置情報408を確認および登定して再配置素を修正や追加や削除などをすることができる。

[0042]上記の処理を行うことによって、取得した 使用状況情報および設定されたグラス関性に基づいて、 ストレージサブシステム200において韓型領域の物理 的再配置を自動的に行い、ストレージサブシステム20 0の最適化を行うことができる。さらに上記の再配置対 断および実行の処理を繰り返して配置を修正していく とによって、使用状況の変動やその他の最適化認差要因を安促していくことができる。

[0043]特に、上記の処理により、ユーザまたは保守具は再配置による最適化を簡便に行うことができる。ユーザまたは保守員は、記憶数置500をクラス600という単位で管理できるため、記憶装置500位能やついて管理する必要はない。さらに、ユーザまたは保守日におい、必要に応じて同一の属性を持つクラス600を設定して、1つの管理単位として扱うことができる。ただし、1つの配慮装置50が1つのクラス600を構成すると見なして、1つの記憶装置50が1つのクラス600を構成すると見なして、1つが記憶装置50が1つのクラス600を構成すると見なして、1つが記憶装置50が1つのカラス600を構成すると見なして、1つが記憶装置50が1つのカラス600を構成すると見なして、1つが記憶装置50が1つのカラス600を構成すると見なして、1つが記憶装置50が1つのカラス600を構成すると見なして、1つが10を指したして上記の再記をの必要を行うことも可能である。

0で行われる処理 (ジョブ) の特徴やスケジュールを考慮して、上記の再配置を自動的に行うことができる。一般に、計算機システムで行われる処理と、この処理に伴う1/Oは、コーザによって作成されたスケジュールに関って行われる。ユーザは、特に最適化の対象としたい処理を有する場合、処理の期間を特定することが可能であり、本実施の形態で説明した再配置り処理によって、コーザは関心のある期間を指定して再配置判断の処理をストレージシステム200に行わせ、すなわち、前記期間の使用状況情報に基づいて上記の再配置による最適化を実現することができる。また、計算機システムで行われる処理はよび1/Oの傾向は目毎、月毎、年毎などの周期性を示す場合も多い。特に、処理が定型業務に基づく処理である場合には、周期性が配着となる。前述の場

合と同様にユーザは、周期において特に最適化対象として関心のある期間を指定して再配置による最適化を行うことができる。また、再配置実行処理313では、ストレージンステム200内で格納内容のコピーを伴うが、35 ユーザはストレージンステム200があまり使用されていない時刻やホスト100で実行されている処理の要求 処理性能が低い期間を再配置実行処理313の実行時刻として設定することで、ホスト100での要求処理性能が続い処理のストレージンステム200への1/Oがコ

[0045] なお、記憶装置500は、それぞれ異なる性能、信頼性、特性や異性を持っていてよく、特に具体的には、斑気ディスク装置、斑気デーブ装配、半導体メャリ (キャッシュ) のように異なる記憶媒体であっても15 よい。また、上記の例では未使用関媒情報407は物理関域に基づいて記述されているとしたが、未使用の物理関域に対応する論理領域(議理アドレス)に基づいて記述されているよい。

ピーにより阻害されることを回避できる。

[0046] <第二の実施の形態>本実施の形態では、 20 使用状況情報としてのディスク装置使用率の適用と、ク ラス600の上限値およびクラス600間の性能順位に よる再配置判断について説明する。

[0047] 図10は、本発明の第2の実施の形態における計算機システムの構成図である。

10048] 本実施の形像の計算機システムは、ホスト10048] 本実施の形像の計算機システムは、ホスト100、ディスケアイシステム201、制御総末700を有してたる。本実施の形像における計算機システムは、第1の実施の形像でのストレージサブシステム200をディスクアレイシステム201とし、記憶装履500をパリティグルーブ501としたものに相当する。

[0049] ディスクアレインステム2011は、創御的300とディスク装配502を右する。 物御的300 は、第10実施の形態での制御的300に相当する。 ティンギャラ

1D (ディスクアレイ) を構成しており、このn台のデ イスク装置502は、n台(nは2以上の整数)でRA データが残りの1台に格納されるといった冗長性上の関 タを含めた格納内容が、並列動作性向上のためにn台の ディスク装置502に分散格納されるなど、データ格納 1を動作上の1単位とみなすことができるが、冗長性や イスク装置502による組をパリティグループ501と 台のディスク装置502の格納内容から生成される冗長 上の関係を持つ。この関係から各パリティグループ50 台数nなどにより実現するためのコストや性値特性など 呼ぶ。RAIDの性質として、1つのパリティグループ 501に含まれるn台のディスク装置502は、n-1 が異なるため、ディスクアレイシステム201を構成す るにあたって、レベルや白数nの異なるアレイ(バリテ 係を持つ。またn台のディスク装置502は、冗長デ 45

イグループ 5 0 1)を混在させることも多く、またパリ ティグループ 5 0 1を構成するディスク装置 5 0 2 につ

ය

いても、性能や容品などによりコストが異なるため、ディスクアレイシステム201を構成するにあたって設めなコストパフォーマンスを実現するために性能や容配の知なる複数簡のディスク装図502を用いることもある。よって本実施の形態においてディスクアレイシステム201を構築する各パリティグルーブ501は性能、信頼性、特性などの属性が同一であるとは限らず、特に性能について差異があるとする。

[0050] 本実施の形態における論理/物理対応情報 400の一例を図11に示す。

[0051] 温度アドレスは、ホスト100がリード/ライト処理310で用いる鍵型的液を示すアドレスである。物理アドレスは実際にデータと前記元長データが格得されるデイスク装配502上の領域を示すアドレスであり、パリティグループ番号と各々のディスク装配番号およびディスク装配がアドレスからなる。パリティグループ番号と各々のディスク装配番号およびディスク装配がドレスが多く。初野部300は、RA1Dの動作として、冗長データに関する情報を前記リード/ライト処理310などで削いて処理するが、本実施の形態の説明では、パリティグルーブ502を動作上の1単位として、記録するだか、前記処理に関してはここでは特にふれ

[0052] さらに第1の実施の形態と同様に、パリティグループ501は、ユーザによってまたは初期状態として複数の組(クラス600)に分類されており、クラス600に分類はクラス600に示す。 クラス構成情報401に設定されている。クラス構成情報401に設定されている。クラス構成情報401の一個を図12に示す。 [0053] クラス帯号は各クラス600に属するパリティグループ数は各クラス600に属するパリティグループが発売がリティグループ部号はカクラス600に属するパリティグループが発売を示す。パリティグループ部号はカクラス600に属するパリティグループ部号においる。本実施の形態におけるクラス 届性情報402に設定されている。本実施の形態におけるクラス 属性情報402の一例を図13に示す。

(0054) クラス番号は、各クラス600を示す番号である。使用非上限値は後述のディスク使用単の許容範囲を示す上限値であり、クラス600の属するパリティグループ501に適用する。クラス同性能域位は、クラス600間の性能類位(後全の小さいものが高性能とする)である。クラス間性能質位は各クラス60を構成するパリティグループ501の前述の性能差異に基づく。所配置実行上限値および固定については後述する。「再配置実行上限値および固定については後述する。「再配置実行上限値および固定については後述する。

10056] 卸卸部300は、第1の実施の形態と同様に、リード/ライト処理310において使用したディスク を表記502の使用時間を取得して単位時間当たりの使

用時間(使用率)を求め、さらに、ディスク装置502 が属するベリティグループ501について、使用率の平均を算出し(ステップ1300)、使用率平均を、リード/ライト対象となった論理領域についてのディスク装 05 置使用率として論理領域提用状況情報403に記録する(ステップ1310)。また制御部30は、パリティグループ501に対応する全論理領域のディスク装置使用率の指を求め(ステップ1320)、パリティグループ501の使用率として物理領域使用状況情報404に

[0057] 本実施の形態における論理節域使用状況情報403および物理領域使用状況情報403および物理領域使用状況情報404の一例を図

15 および図16 に示す。 [0058] 日時はサンプリング問隔 (一定期間) 毎の 15 日時を示し、論理アドレスは陰壁倒域を示し、パリティ グループ番号は個々のパリティグループを示し、韓壁倒 域のディスク装置使用率およびパリティグルーブ使用率 はそれぞれ前記サンプリング回隔での平均使用率を示

す。上記のようなディスク装置502の使用率はディス)ク装置502にかが5負荷を示す値であり、使用率が大きい場合は、ディスク装置502が性能ボトルネックとなっている可能性があるため、再配置処理で使用率を下げることによりディスクアレイシステム201の性能向上が期待できる。

25 【0059】次に、再配置判断処理312について図17で数明する。

【0060】制御邸300は、各クラス600について、クラス600に属するパリティグループ501をク

ス株成債額401か590年75 (ステップ130)
 300)。 様いて、初卸節301、第10実施の形態と同様の再配置判断対象期間情額405を参照して対象期間を取免し、きちにパリティグルーブ501について、対象期間の勢理関域使用状況情報404のパリティグルーブ使用率を取得し、集計する (ステップ1320)。 様いて、初週部300は、クラス属性情報4040ペッを転用、

を取得し、さらにパリティグループ501について、対象期間の物理関域使用状況情報404のパリティグループ使用率を取得し集計する(ステップ1320)。 続い35 て、制御節300は、クラス属性情報402を参照してクラス60の使用率上限値を取得する(ステップ133)の 制御節300は、パリティグループ使用率とクラス上限値より数とは、パリティグループ501の使 用率を減らすために、パリティグループ501の使 用率を減らすために、パリティグループ501に対応する 高強理領域の再配置が必要と判断する(ステップ134

[0061]税いて、制御部300は、対象期間の論理 領域使用状況情報403を参照して、再配置が必要と判 断したパリティグループ501の各物理領域に対応する 論理領域のディスク装置使用率を取得し集計して (ステップ1350)、ディスク装置使用率の混りに は計して (ステップ1350)、ディスク装置使用率の大きいものか 5、再配置する論理領域として違択する (ステップ1360)。 論理領域の違択は、パリティグループ501の

ていき、クラス600の使用率上限値以下になるまで行う (1370)。 ディスク装置使用率の大きい論理領域は、パリティグループ501の使用率に対する影響も大きく、またホスト100からの論理領域に対するアクセス頻度も大きいと考えられるため、ディスク装置使用率の大きい論理領域を優先的に再配置することで、ディスクアレイシステム2010効果的な性能改善が期待でき

| 0062| 制御部300は、選択された論理領域についての再起置法となる物理領域を探す。制御部300は、クラス属性情報402を参照し、パリティグループ501の属するクラス600より性能原位が高位のクラス600(高性能クラス)に注目し、クラス構成情報401および第1の実施の形像と同様の未使用領域情報407を参照して着性能クラスに属するパリティグループ501の未使用物理領域を取得する(ステップ138

【0063】さらに、制御部300は、各未使用物理函域について、再配置先とした場合のパリティグループ使用却の干弱値を求め(ステップ1390)、未使用物理倒域の中から、再配置先とした場合に高性能クラスに設定されている上環値を超えないと干剤できる未使用物理領域を、再配置先の物理領域として選択し(ステップ1400)、選択結果を第1の実施の形態と同様に、再配置情報408に出力する(ステップ1410)。選択した全ての論理領域について再配置先の物理領域を選択した全ての論理領域について再配置先の物理領域を選択した金にの過程の域について再配置先の物理領域を選択した金にの過程の域を選択し

[0064] 本実施の形態において、制御的300は、第1の実施の形態に加えてベリティグループ情報409を保存し、ベリティグループ情報409、諸理関域使用状況情報403、及び物理関域使用状況情報404から使用非改賞を算出する。

[0065]パリティグルーブ情報409の一例を図18に示す。パリティグルーブ番号は個々のパリティグルーブ501を示す番号である。RAID構成はパリティグルーブ501が構成するRAIDのレベルやディスク台数や冗長度構成を示す。ディスク装置性能はパリティグルーブ501を構成するディスク装置を能はパリティグルーブ501を構成するディスク装置を配配においてディスク装置使用卵の大きい論理領域の再配置先を高性能クラスのパリティグルーブ501とすることで、同一食荷に対するディスク装置使用時間を超縮でき、論理領域の再配置後のディスク装置使用時間を超縮でき、論理領域の再配置後のディスク装置使用時間を超縮でき、論理領域の再配置後のディスク装置使用時間を超縮でき、論理

[0066] 再配置実行処理313は、第1の実施の形態と同様に行われるが、図19に示すように、制御断300は、再配置のためのコピーを行う前にクラス属性情報402を参照し、再配置元および再配置先のクラス60について、ユーザによってまたは初期条件として設定された再配置実行上限値を取得する(ステップ1500)。さらに物理領域使用状況情報404を参照して、

再配置元および再配因先のパリティグルーブ501の直近のパリティグルーブ使用串を取得し (ステップ1510)、比較の結果少なくとも一方のクラス600においてパリティグルーブ使用串が再配置実行上限値を超えていり、た場合は (ステップ1520、1530)、再配置実行物理313を中止または延期する (ステップ154

[0067]上記処理によりユーザは、パリティグループ501の使用率が大きくすなわち負荷が高い場合に向いるコピーによりさらに負荷が生じることを回避することができ、また回避のための上限値をクラス600位に任意に設定することができる。

• 6 [0068]上記のように処理することによって、ディスク装置502の使用状況に基づいて物理的に再配すする論理領域の選択、および再配置先の物理領域の選択。

を、クラス構成および属性に基づいて行い、再配置によりディスク装置502の負荷を分散して、各クラス600に属のに数定されている使用率上限値を、クラス600に属するパリティグループ501の使用率が組えない配置をするパリティグループ501の使用率が組えない配置を収集することができる。さらに再配度判断および実行の処理を繰り返して配置を修正していくことができる。目の69]再配置判断処理312において、毎期約30変動や予測級差を吸収していくことができる。

00は、対象期間の物理領域使用状況情報404のバリ 25 ティグループ使用車や、随理領域使用状況情報403の 路環関域のディスク装置使用車を参照して抵計し、判断 に用いるとしたが、例えば、対象期間の全ての値の平均 を用いる代わりに、対象期間中の上位血個の値を用いる 方法も考えられ、また上位血番目の値を用いる方法も考 30 えられる(mは1以上の整数)。これらの方法をコーザ が選択できるようにすることで、ユーザは使用状況の特 域的な部分のみを選択して用い、再配置判断処理312 を行わせることができる。

610470]上記の再配理判断処理312において、耐 10070]上記の再配理判断処理312において、耐 10070]上記の子レイシスケム2010全ての 7ラス600について、雑理関係の所配置の必要なバリ ティグルーブ501の検出を行うとしたが、前記検出の 前に制御部300かクラス層体指領402を参照し、固 定属性が設定されているクラス600については、検出 10 の対象外としてもよい。また同様に、網線節300が、 リティグルーブ情報409を参照し、固定風柱が設定されているバリティグルーブ方 れているバリティグルーブ501については検出の対象 外としてもよい。また、再配置判断処理312におい

て、制御部300は、高性能クラスに属するパリティグ45 ループ501の未使用物理関係から再配置先の物理関域を避保するとしたが、固定属性が設定されているクラス600については対象外として、さらに性能質位が高位のクラス600を高性能クラスとして扱うようにしてもよい。また固定属性が設定されているパリティグループ5050にには対象外としてもよい。上記のように固

使用率から選択した論理領域のディスク使用率を減算し

ය

定的性が設定されているクラス600またはパリティグループ501を扱うことによって、ユーザは上記の自動的な再配原の駆撃を生じさせたくないクラス600またはパリティグループ501を設定し、再配配の対象外とすることができる。

(0071) < 第三の実施の形態>本実施の形態では、 同一クラス600 件での再配質判断について設明する。 本実施の形態での計算機システムは、第2の実施の形態 と同時である。ただし、本実施の形態では1つのクラス 600に複数のパリティグルーブ501が属する。本実 施の形態での処理は、再配置判断処理312を除いては 第2の実施の形態と同様である。また、再配質判断処理 312についても、再配置する。また、再配質判断処理 716001は、第2の実施の形態と同様である。

/ 15 U U) 17、3/2の美婦の形態と同様である。 【0 0 7 2】本実施の形態での再配置判断処理3 1 2 における、in配置光の物理関係の選択について図2 0 で設

位が高位のクラス600から選択するが、本実施の形態 プ501から選択する。制御邸300は、クラス構成情 2600に属する再配置元以外のパリティグループ50 月師部300は、各末使用物理領域について、再配置先 (ステップ1620) 、未使用物理質域の中から、再配 限値を超えないと予測できる末使用物理領域を、再配置 先の物理領域として溢択し (ステップ1630)、 選択 **情果を第2の実施の形態同様に、再配置情報408に出** 域について再配置先の物理関域を選択し終えたら処理を 【0073】 第2の実施の形態では再配配先の物理領域 を可配置元の物理領域の属するクラス600より性能観 徴401と米使用領域情報407を参照して、同一クラ 1の米使用物理菌域を取得する (ステップ1610)。 **页先とした場合に同一クラス600に設定されている上** カする(ステップ1640)。 再配置する全ての論理領 では同一クラス600の再配置元以外のパリティグルー とした場合のパリティグループ使用率の予測値を求め 終了する (ステップ1650)。

10074]上記の処理により、同一クラス600内においてディスク装置502の負荷を分散することができる。上記の処理方法は例えばディスクアレイシステム201のペリティグループ501が全て1つのクラス60 (単一クラス)に属する情域に適用することができる。また、例えば、第2の実施の形態で設明した処理方法と組み合わせた場合に、再配置先の未使用物理領域の選択において、再配置元のクラス600より性結婚値が高値のクラス600に適当な未使用物理関域が得られなかった場合や、性能断値が最上位のクラス600での処理に適用できる。第2の実施の形態で説明した処理方法と組み合わせた場合は、第2の実施の形態での処理方法と対合うフス600について異なる使用車上限値を用いてもよく、すなわち、その

二種類の使用率上限値または差分を有してもよい。

存眠2001-67187

【0075】く第四の実施の形態ン本実施の形態では、第2の実施の形態での再配置判断処理312において、再配置元のクラス600より性能順位が高位のクラス600より性能順位が高位のクラス600(高性能クラス)に再配置先の未使用物理領域が見つからなかった場合に、再配置先を得るために先立って行われる、性能順位がより低位のクラス600(低性能クラス)への高性能クラスからの再配置の処理についても8mm+x

0 [0076]本実施の形態での計算機システムは、第2の実施の形態と同様である。本実施の形態における再配置判断処理312について図21で説明する。

[0077] 制御部300は、高性能クラスに属するバリティグループ501をクラス構成情報401から取得15 する(ステップ1700)。 続いて制御部300は、第10実施の形態と同様の再配置判断対象期間情報405を参照して対象期間を取得して対象期間を取得(ステップ1710)、対象期間の論理関鍵使用状況情報403を参照して、バリティグループ501の各物理関域に対応する諸理関域のディスク装置使用単を取得してステップ1720)、ディスク装置使用単の小さいものから、低性能クラスへ再配置する論理関域として過択する(ステップ1720)、ディスク装置使用単の小さいものから、低性能クラスへ再配置する論理関域の選択は必要なだけ行われる(ステップ1740)。このとき論理関域の選択は必要なだけ行われる(ステップ1740)。

25 [0078] 続いて制御師300は、選択された韓理領域についての再配置先となる物理領域を、低性能クラスに属するパリティグループ501から選択するが、再配置先の物理領域選択の処理は、第2の実施の形態での処理設明において再配置先としている高性能クラスを低性現象ラスと競み替えれば、第2の実施の形態での処理と同様である(ステップ1750)。また、本実施の形態におけるその他の処理も第2の実施の形態での処理と同様である(ステップ1750)。また、本実施の形態におけるその他の処理も第2の実施の形態での処理と同様である。

[0079]上記の処理を行うことで、第2の実施の形態での再配置判断処理312において高性能クラスに再配置先の未使用物理領域が見つからなかった場合に、高性能クラスから低性能クラスへ論理領域の再配置を、高性能クラスへの再配置に先立って行い、再配置先の未使用物理領域を高性能クラスに用意することができる。制御第300は、上記の処理を必要に応じ繰り返し行っ

て、十分な未使用物理領域を用意することができる。 [0080] 臨理領域の再配置先を低性能クラスのペリティグループ501とするため、同一負荷に対するデイスク使用時間が再配置について増大し、強理領域の再配 (配後のディスク使用率の小さい強建領域から再配置していくようにすることで、増大の影響を最小限に抑えることがで

[0081] < 第五の実施の形倣>本実施の形倣では、 50 クラス600の属性の1つにアクセス種別属性を設け、

アクセス権別属性を用いてシーケンシャルアクセスが顕着に行われる論理領域とランダムアクセスが顕著に行われる論理領域と多、他のパリティグループ501に自動的に参理的再配置して分離するための再配置判断につい

[0082] 本実施の形態における計算機システムは図10に示したものである。本実施の形態では、第2の実 第の形態での説明に加え、制御的300が保持する下記の情報を用いる。

[0083]本実施の形態でのクラス属性情報402の一般を図22に示す。この例では、第2の実施の形態での例に対しアクセス種別が加えられており、クラス600アクセス種別が、例えばシーケンシャルに設定されている場合は、クラス600がシーケンシャルに設定されている場合は、クラス600がシーケンシャルアクセスに好適であると設定されていることを示す。

【0084】本実施の形態での論理領域使用状況情報403の一例を図23に示す。この例では、第2の実施の 形態での例に対し、シーケンシャルアクセス単およびテングムアクセス単およびテングムアクセス単お加えられている。 [0085] さらに、本実施の形態において制御部300は、第2の実施の形態に加え、アクセス艦別基準値情報410と論理領域属性情報411を保存する。 [0086] アクセス艦別基準値情報 410の一倒を図

24に示す。ユーザによりまたは初期条件として、アクセス種別基準値情報410には後述のアクセス種別の判定に用いる基準値が設定されている。また、路理領域属性積和110一向各図25に示す。アクセス権別とソトは、各路理領域について顕著に行われると期待できるアクセス種別であり、ユーザが設定する。固定については条法する。

【0087】本実施の形態での処理は、使用状況情報取得処理311および再配配判断処理312を除いては第二の実施の形態と同様である。

[0088] 本実施の形態における使用状況情報取得処理311について図26で説明ナる。

[0089] 制御部300は、第20実施の形態での使用状況情報取得処理311と同様に、韓理確確についてのディスク装置使用率を算出し(ステップ1800、1810)、リード/ライト処理310での使用単内容を分析して、使用率についてシーケンシャルアクセスとランダムアクセスの比率を算出し(ステップ1820)、使用率およびアクセス種別比率を確理領域使用状況情報403のは、第2の実施の形態と同様にバリティグループ使用率の算出と物理領域使用状況情報404への記録を行う(ステップ1840、1850)。

[0090]本実施の形態における再配置対断処理312において、再配置する協理保険の選択は終20実施の形態と回接である (ステップ1990)。再配置判断処理312での再配置先の物理領域の選択について図27

で説明する。

【0091】制御節300は、韓理衛域使用情報403を参照し、再配置する韓國域についてのシーケンシャルアクセス率を取得し、ステップ1910)、アクセスの 種別基準値指標410に設定されている基準値と比較する (ステップ1920)。ソーケンシャルアクセス率が 基準値より大きい場合、制御節300は、クラス原性前 権402を参照し、アクセス種別がシーケンシャルと設定されているクラス60(ソーケンシャルクラス)が 10 存在するが調べる (ステップ1950)。シーケンシャルクラスが存在する場合、制御節300は、グラス構成情報401と未使用領域情報401を参照して、シーケンシャルクラスに属する再型面に以外のパリティグルーブ501の未使用領域情報407を参照して、シーケンシャルクラスに属する再型面に以外のパリティグルーブ501の未使用物理例域を取得する(ステップ196)。

15 0)。 さらに制御節300は、各未使用物理阅域について、再配置先とした場合のパリティグルーブ使用非の子 調値を求め (ステップ1970)、未使用物理関域の中から、再配置先とした場合にシーケンシャルクラスに設定されている上限値を超えないと予測できる未使用物理 20 領域を、再配置先の物理関域として選択し (ステップ1980)、選択結果を第2の実施の形態同様に再配置情 報408に出力する (ステップ1990)。制御節300は、使用率予測値を、第2の実施の形態に同様のパリティグルーブ情報403および物理阅域使用状況情報402を実施の形態における論理函25 域使用状況情報403および物理阅域使用状況情報40

[0092] 前記の比較において、ツーケンツェルアグ
れス華が基準値以下である場合、影響節300は、諸理 衛塚風性情報411を参照し、諸理協成についてアクセ ス層別とントがソーケンシャルと設定されているか認く る (ステップ1940)。アクセス観別とフトにソーケ ソシャルと設定されていた場合、上記と同様に制御第3 00は、シーケンシャルクラスの有無を強く(ステップ 1950)、シーケンシャルクラスが存在する場合は、

シーケンシャルクラスから再配配先の物理領域を選択する(ステップ1960~1990)。【0093】前記の比較において、シーケンシャルアクセス率が前記基準値以下であり、さらにアクセス圏別に

10093] 而記の比較において、シーケンシャルアクセス年が前記基準値以下であり、さらにアクセス層別にントがシーケンシャルでなかった場合、またはシーケンシャルクラスが存在しなかった場合、制御網300は、 第20実施の形態と同様に、シーケンシャルクラス以外のクラス500から再起図先の物理関係を選択する(ス

テップ2000)。 [0094] 上記の処理により、同一パリティグループ 45 501での顕著なシーケンシャルアクセスとランダムア クセスの確在に対し、各クラス600に属性として設定 されたアクセス極別と使用単上限値を用いて、シーケン シャルアクセスが顕著に行われる論理領域とランダムア クセスが顕著に行われる論理領域とを、異なるパリティ 50 グループ501に自動的に再配置して分離、すなわち異

ためにクラス 異性情報 402 が各クラス600 について

なろディスク装置502に分離することができ、特にラ ングムアクセスに対する応答性能を改善することができ

[0095]また、上記の処理においては制御部300 は、シーケンシャルアクセスに注目して再配置による自 動的分離を行うとしたが、同様にランダムアクセスに注 目して前記分離を行うことも可能である。

[0096] 上記の再配配判断処理312において、再 県領域属性情報411を参照し、論理領域に固定属性が 城がある場合、固定風性を設定することで論理領域を再 配置の対象外とすることができる。上記の固定異性に関 する処理は論理領域属性情報411を用いることで、前 配置する論理節域を選択した時点で、制御部300が輪 **指定されている場合は、論理即域を再配置しないとすれ** ば、ユーザが特に再配置を行いたくないと考える論理領 むの実施の形態にも適用できる。

[発明の効果] ストレージサブシステムのユーザ、また 1保守員が、記憶領域の物理的再配置による配置最適化 を行うための作数を筋優にすることができる。 [図面の簡単な説明] [0097]

[図1] 本発明の第1の実施の形態での計算機システム の構成図である。 [図2] 本発明の第1の実施の形態でのリードノライト 処理310および使用状況情報取得処理311のフロー

[図3] 本発明の第1の実施の形態での論理/物理対応 チャートである。

【図4】本発明の第1の実施の形態での再配置判断処理 312のフローチャートである。

14位400の一例を示す図である。

[図5] 本発明の第1の実施の形態での再配置判断対象 [図6] 本発明の第1の実施の形態での再配置情報40 引間情報405の一般を示す図である。

[図7] 本発明の第1の実施の形態での未使用領域情報 8の一例を示す図である。

[図8] 本発明の第1の実施の形態での再配置実行処理

107の一倒を示す図である。

[図9] 本発明の第1の実施の形態での再配置実行時刻 313のフローチャートである

[図10] 本発明の第2の実施の形態および第五の実施 言語4060一度を示す図がある。

[図11] 本発明の第2の実施の形態での論理/物理対 の形態の計算機システムの構成図である 応位数4000−8を示す図である

[図13] 本発明の第2の実施の形態でのクラス属性情 四4010-因を示す図である。

[図12] 本発明の第2の実施の形飾でのクラス構成情

M102の一例を示す図である。

【図14】 本発明の第2の実施の形態での使用状配情報 取得処理311のフローチャートである。

[図15] 本発明の第2の実施の形態での論理領域使用 状況情報403の一例を示す図である。

【図16】本発明の第2の実施の形態での物理領域使用 状況情報404の一例を示す図である

【図18】 本発明の第2の実施の形態でのパリティグル 【図17】本発明の第2の実施の形態での再配置判断処 埋312のフローチャートである。 S

【図19】 本発明の第2の実施の形態での再配置実行処 ープ情報409の一個を示す図である。

埋313のフローチャートである

으

【図20】本発明の第3の実施の形態での再配置判断処 【図21】本発明の第4の実施の形態での再配置判断処 母312のフローチャートである。

吐312のフローチャートである。

【図22】本発明の第5の実施の形態でのクラス属性情

12

[図23] 本発明の第5の実施の形態での論理領域使用 ₩402の──既を示す図である。

伏妃情報403の一例を示す図である。

【図24】本発明の第5の実施の形態でのアクセス種別

格学値信報410の一例を示す図である。

2

[図25] 本発明の第5の実施の形態での論理領域属性

[図26] 本発明の第5の実施の形態での使用状況情報 前悔4110一回を示す図わせる。

取得処理311のフローチャートである。

【図27】 本発明の第5の実施の形態での再配置判断処 吐312のフローチャートである。

23

[作号の説明]

100 #21

200 ストレージサブシステム

201 ディスクアレイシステム

8

リード/ライト処理

強領 300 使用状况情報取得処理 311

310

再配置判断処理 312

再配置实行処理 313

33

福里/物風対応位報 クラス構成情報 401

クラス異性情報 402

論理領域使用状況情報 403 物理領域使用状況情報

404

再配置判断对象期間信報 再配置实行時刻情報 405 406

未使用領城情報 407 再配置情報 409 408

パリティグループ情報

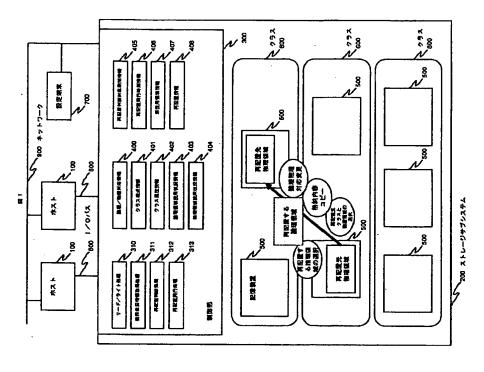
42

アクセス種別基準値情報 論理領域属性情報 4 1 0

記憶装置 411 500 パリティグループ ディスク装置 501 502

900 ネットワーク 800 1/0/3

[<u>図</u>]



[6図]

[図24]

å

824

1998年8月11日 22時0分 (国第73十) 日第

> 日時更新情報 世

アクセス和政治事業(%

S

[図2]

制御部は路場信はのアドレスを物理信仰のアドレスに食物 ホストが経過音楽を密定してリード/ウイトを東水 制の部はリード身は物理保証からホストへ、 bイト時はホストから物理機能へデータを転送 金罗郡江路国信其使用女民所集长光新 复穿着订告运输法使把权权际电力判据 会がなどの名 リードノライトの種

[图3]

8

		教理アドレス
1447 F C C	記憶装置番号	記憶装置のアドレス
886~0	0	666~0
1000-1989	0	1000-1888
2000-2999	1	666~0
3000~3889	1	1000~1999
		3

[図5]

199948A118 8B30#	1099年8月11日 17時16分	毎日(十24時間)	
斯 结 600	株7日時	和的英統領	

再配置対象部的参信の対象別数を次回分に更新 再記書先のクラスを選択 再記書先の未使用権理領域を選択 英記書を行う人を禁忌祭業を設定 **会评的以他国际政策用状况未参照** 紀式指示中耳氏的反射に対力 [34] クラス異性情報を参加 クラス異性情報を参照 -[=

[図6]

		Ħ	再配置元物電信號	製菓	再配置先他理樣媒
\$	***************************************	68 李松斯四	記憶器配件 アドレス	648 85527	記憶発電点 アドレス
-	0-8-0	٥	0~000	10	000~0
2	1000~1989	0	1000-1888	0 1	1000-1999

[図7]

Ħ

記憶院定書号	記憶物気内アドレス	使用/未使用
0	666~0	铁用
0	1000-1996	WW.
0	8682-0002	未使用
0	8686~0006	世級者

[図10]

[88]

朱使用菜类情報表,再因配先物理或基在使用11、再因避先物理链域老未使用17黄则 英国ノ他国党の役割を其的最先を延縮をから和別的会議を対す。 気がむな年の間元物を投がの物を内容を本典の関先を基礎をにコピー 2

再配置発行時間信仰の日時を次回分に更新

- [

[311]

=

17.2					
			4-4	•	九五千一9
	1074774-7 84	6-6 2812	とつけて 内部製造	200 SER.	記録数額内 アドレス
665~0	100	0	868~0	20	0~0
1000-1989	100	۰	1000-1999	20	1000~1888
2000~2999	101	-	. 088~0	4.1	0.66-0
3000-3999	101	1	1000-1999	÷	1000-1988

[🖾 1 2]

12

パリティブループ音号 100、110、120 101、111 102、112、122、132	3. 3. 2. 4. 4. 4.	1
102, 112, 1	•	
101. 1	2	-1
100. 110.	3.	
パリティグループ	パリティグループ数	-

S 201 #4 X 9 T L 4 5 X # L 0 0 0 7/3" とールカトテレン

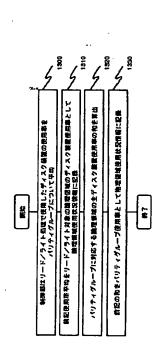
[國13]

BE 13

(金)	,	田	1	
再配體実行上函数(%)	7.0	90	0.6	
クラス間性他派位	1	2	3	
使用卷上提供 (%)	. 09	7.0	80	
99X89	۰	-	~	

[図14]

4



[図15]

5

10000000000000000000000000000000000000	6年7ドレス 0~688 1000~1889	ディスク教徒使用性(M) 1.8 3.2
1980#8A118 8#18A	1000-100	30
Social Birkshops	1000~1999	44 M

[2] 1 6]

-

69	パリティグループ番号	使用平 (地)
	100	89
1998年8月11日 日降0分	101	5.2
	001	7.0
1000年8月11日 8年15分	101	9
	001	2.4
1989年8月1日 日本10分	101	9.7

[图18]

<u>.</u>

_	_			_
¥	1	展	1	
ディスク装置性能	110	100	9.0	
RAIDORE	RAIDS 3D1P	RAID1 1D1P	RAIDS SOIP	
パリティグループ教与	100	101	102	

[图22]

822

99X84	号 佐州年上版館 (%)	クラス間 性線照位	有配置 実作上限数 (%)	X 8	19セス国別
°	60	1	0.4	-	-
-	7.0	8	0.8	_	_
2	0.8	ę.	0.6	1	シャンジャル

[2] -

パリティグループ使用率がクラスの使用率に開始とり大きしていティグループは再定式の必要があると対断する 何同部は各クラスについて、馬するパリティグループを取得 パリティグループが属するクラスの使用率上級包含取得 対象規関のパリティグループ使用命を取得、進計 異の個性語の対象は四年表現

解的物は再配配元および再配配先のクラスの再配置実行上限位を取得

2

[図19]

. E

再配置元および再配置先のパリティグループの使用事を取得

サイスン会議機能では登回の大中の協議機能がある。 親間終する間機器終して経過 アド 再記載の必要なパリナイグループについて、対応する主動機関項のディスク設置 使用字を対象規則について政制・条件

お記令本後用物理会はについて、お記型先とした場合のパリティグループ使用等 を**そ**よ 制御群は高性能クラスの水使用物を循絡を改得 過収分を延伸して 使用率上回復以下になるか

中土東仁は田州

4 8

A配置名について ペリティグルーブ保証部が クラスの開配電製作上配置を 超えているか

1520

原記器だについた パリチェグループ製配料が クラスの製配製製作上報告件 超光でいるが

對方部以非記憶元物協議的形成內容を再記記を他認動以にコピー 路間/他国均乙仿總を民選、未使用指域情報を表現、再提供政行等対抗報告更新

が記予機関が実施なクラスの使用等上側質を超えない余使用物面質量を 専門製作の管理器域として過収 退稅條果を再配置債権に出力

体配数する会替制を設について を担信性を研究したか

[図23]

2 3

*	加着アドレス	ディスク経星 (使用率 (%)	シーケンシャル アクセス章 (%)	ランダム アクセス母 (%)
	666~0	18	7.8	2 2
· 1998年8月11日 8時0分	1000~1989	3.2	5.2	4 8
	886~0	80	0 8	80
1998年8月11日 8815年	1000-1888	3.0	08	o,
	666~0	2.2	8.8	1.8
1888年6月11日 8時30余	1000-1999	8.8	8 7	2 9

. .

特開2001-67187

[図20]

0 2 0

群記名ネ使用物理病域について、再配置先とした場合のパリティグループ使用率 を予測 你記字選集が収配両一クラスの使用事上限度を超えない未使用物理協権を 再配置をの他環境域として過収 前部物は其配理元パリティグループの第一クラスの、其限部元パリティ グループ以れのパリティグループの条件用を直接基を取得 異形置する処理領域の退択(複数選択されうる) 因代略與各种配數價值に出力 東記載する金銭は保護について 物理領域を設択したか 2 [=

[221]

2

制御的は実性他クラスに属するパリティグループを取得 再記憶和新の対象和図を取得 g -

資格権クラスに属するパリティグルーグについて、対応する全部通過機の ディスク統型使用学を対象が関について教育、集計 ディスク使用事業計算長の小さな間理機関から、 在性費クラスへ再配置する簡単倒域として選択

進択した設備的項に対する、信性数クラスモの有限数先の物理資格の過失 施欠した機能保証により 必要な信息を指揮できるか

[図25]

2 2

物理アドレス	オペコ経験とみると	¥
866~0	_	_
1000~1000	-	-
2000~2888	シーケンシャル	
3000~3000	1	¥B

[図26]

9

作品ディスク技会を用されたび作品ディスクは登役用写についての シーケンシャルアクセスとランダムアクセスの比率を終び監督は原外交付書に記録 **な別係所等がなかしードノウイトが彼の間凹回向のディスク技術仮用等とする** パリチィグループに対応する智慧物質の会ディスク技術的用者の他を禁止 数的の智やパッチィグルーン数形形として他動図基数形式の数量に記載 制御部はリード/ライト処理で使用したディスク独画の使用事を パリティグループについて平均 前記ディスク装置使用率のアクセス短割比率を算出

14

[図27]

B27

Fターム(参考) 5B065 BA01 CA30 CC01 CC03 EK01 5B082 CA11

> **製造的はソーケンシャケケルスの、第四部代ペンサ・** グペープ以近のパリティグループの米数配割延延基本数据 **貸配合来供用物質信仰について、発売配金した場合の パリティグループ製用等を予測** 算配予済値が前応用ークラスの使用率上機能を燃えない。 本使用物理関性を再配置たの物理応はとして設別 マーケンフトラクロスは存在するか 過氧結果を再配置者無に因力 異記載する雑類信仰の溢択(複数温飲されうる) 食物をは続次した物質医薬の シーケンシャルアクセス手を取得 シーケンシャルクラス以外からの 再配置先物理保護の選択 制改都は選択した情理協議の アクセス領別にントを取得 ツーケンシャルアクセス都は 基準数より大きいか アクセス部宮ドントは
> ツーケンツャルや の配置する金額は位列につい 他は合意を過収したか 数图

フロントページの概念

(72) 発明者 山神 裏司

种奈川県川崎市麻生区王禅寺1099番地 株 式会社日立製作所システム開発研究所内

(72)発明者 荒井 弘治

神奈川県小田原市国府津2880番地 株式会社日立製作所ストレージシステム事業部内

高機能ディスクにおけるアクセスプランを用いた プリフェッチ機構に関する評価

向井 景洋, 根本 利弘, 喜連川 優

東京大学 生産技術研究所

〒 106-8558 東京都港区六本木 7-22-1 Tel 03-3402-6231 Fax 03-3479-1706

E-mail: {mukai,nemoto,kitsure}@tkl.iis.u-tokyo.ac.jp

以作、様々な研究機関で、インテリジェントなディスクに関する研究が行われている。これらの研究で 半導体技術の進步により、プロセッサ、メモリは急速に高性能化しているにも関わらず、コストは急激 **投深されているディスクは、データベースアプリケーションの性能を大幅に向上させる一方で、ホスト** コードの変更が少ないことが特徴として挙げられる。このディスクに於いて、アクセスプランを用いた に下がってきている。このような状況を背景に、単なる容量としてのディスクではなく、高性能のコン 幼中的な1/0 処理を行う、インテリジェントなディスクを機察する。このようなディスクでは、ホスト プリフェッチを行った場合に、どの程度の柱館向上が見込まれるのかを聞べる為に、シミュレーション トローラや大容益のメモリを格扱し、復雄な処理を行うことの出来るディスクに注目が集まっている。 のコードの大幅な変更を必要とする。そこで、本語文では、アプリケーションレベルの知識を理解し、 **火殺を行った。その結果、戈行時間の大幅な削減が期待できることが刺った。**

Evaluation of Prefetching Mechanism Using Access Plan on Intelligent Disk

Kagehiro Mukai, Toshihiro Nemoto and Masaru Kitsuregawa

Institute of Industiral Science, The University of Tokyo 7-22-1, Roppongi, Minato-ku, Tokyo 106-8558, Japan

Abstract

they improve the performance of database applications significantly, dbms code has to be modified a lot. In this paper, we examine the access plan based prefetching approach where the access plan is given to the disk controller. This approach needs very little change in host code. We examine how our access plan based prefetching approach improves the performance in detail. Experiment chips available at cheaper price. Therefore, it has become possible for disks to process complicated tusks. There are several research projects in the field of intelligent disk architecture. Although The progress of semiconductor technologies has made high performance processors and memory results indicate that the execution time decreases dramatically

聞が数ミリ秒であるのに対して、0.1 ミリ秒のオーダー のコントローラには 100MHz のプロセッサが用いられ、 ディスク・キャッシュのサイズはすでに 4MB(殺大 16 MB)[8] に遠している。そのため、ロボットアームなど **遠に高性能化しているにも関わらず、そのコストは急** ー方、データベースの世界では、 DSS やデータ・ウェ 半導体技術の進歩により、プロセッサ、メモリは急 徴に下がってきている。現在のハイエンドなディスク の物理的な動きを必要とするシーク時間や回転待ち時 クの急速な増加に対応する為に、ユーザは膨大な量の ムにおいては、ホストのコンピュータはディスク国の アハウスは急速に大規模化し、システムにおける処理 性能の向上や容量の増加が強く要求されている。デー データを複数のディスクに保存する。 大量なデータを 多数のディスクに保持し、並列にアクセスするシステ ホストとディスクを繋ぐネットワークのバンド値はボ で処理を終えるコントローラの性能が際だっている。 **蜘蝣に大部分の時間を割くことが必要となる。また、** トルネックになりやすい。

以上で述べた事を背景に、単なる容量としてのディ スクではなく、高性館のコントローラや大容量のメモ リを搭載し、複雑な処理を行うことの出来るインテリ ジェントなディスクに注目が集まっている。

し、それを基に効率の良い 1/0 処理を行うインテリジェ 本鑑文では、アプリケーションアベルの凶棍を阻解 ントなディスクを提案する。このようなディスクでは、 アプリケーションフスアの凶猛や阻解できる為、虫荷 が比較的軽い時間を利用し、将来、ホストが参照する もとでプリフェッチを行っておくことにより、キャッ シュヒット郷を向上させることができ、性能の向上が と予想されるデータを効率の良いスケジューリングの 期待できる。

大規模な SMP システムにおいて、スキャン、ハッシュ インテリジェントなディスクに関する研究例として スでデータを処理するディスクレットを用いた研究 [3]. へのデータ転送を削減する研究 [4]、近似データサーチ、 は、データベースにおける SELECT、GROUP BY、 ジョイン、ソートをディスクで行う事により、ホスト ゴリズムを扱った研究 [2]、などが複素されている。こ 外部ソート、データキューブ、画像処理などのアプリ ゲータマイニング、画像処理のスキャンペースのアル ホストのコードの大幅な変更を必要とする。これに対 し、本論文で提案する方式では、ホスト側の変更は少 ケーションに関し、ディスクに内在しストリームベー なく、既存のシステムに適用しやすい事が利点として れらの研究では、複雑な処理をディスク内で行う為、

シミュレーション実験を行った。その結果、大きな実 フェッチが行われた場合、どの程度の性能の向上が見 込まれるのかを評価する為に、データベースのベンチ マークである TPC-H をアブリケーションとして加い、 本額文では、データベースに於けるアクセスプラン が与えられたディスクにおいて、そのアクセスプラン を用い、知的なアクセススケジューリングによるプリ 行時間の削減が烙件できるという結果が得られた。

以下、2章で従来のディスクアクセスについて街中 3章では2章で楚楽したプリフェッチ機構に関する評価 に説明し、問婚点を明らかにする。その後、アクセス プランを用いたプリフェッチ機構について説明をする。 乾燥を行う。 最後に、 5章で、まとめと今後の瞑題を 実験の詳細を述べる。 4章では評価実験の結果を示し、

アクセスブランを用いたディスクブリフェッ

2.1 従来のディスクアクセスとその問題点

Hの関合せの一つである Query3 の SQL と ORACLE9 チャークの一つである TPC-H を考える。まず、 TPC-によって作政されるアクセスプランの一部を図 1に示 アナリケーションの殴として、ゲータベースのヘン

オトは、where 句に聞して、PARTS,LINEITEM,ORDERS のインデックス、またはテーブルにアクセスし、 Nested 図1のアクセスブランに示されるように、 Q8 で、ホ Loops Join を行っている。図1に示したSQL 文に閩 **する ORACLE システムのディスクアクセスについて** 詳しく検討していく。 まず、ORACLE システムは、

の条件を描たすレコード番号を得る為に、PARTS のインデックスにアクセスを行う。その後、 P.TYPE = 'SCONOMY ANOBIZED STEEL' (I)

の条件により、式 (1) より得られたレコード番号の PARTS テーブル、LINEITEM のインデックスにアクセスし、 NESTED LOOPS JOIN を行う。 次に、 P.PARTKEY = L.P.ARTKEY

9 LORDERKEY = O.ORDERKEY

の条件を済たすレコードを見つける為に、ORDERS のインデックス、テーブルに頃次アクセスを行い、式 (2)で得られたレコードと NESTED LOOPS JOIN ODRDERDATE

を行っている。LORDERKEY に関しては、式(2) で



INDEX SCAN | PARTS | P. TYPE - "ECONOMY ANODIZED STEEL" 図 1: (左)Query8 の SQL,(右)Query8 のアクセスプランの一部 L_ORDERKEY = 0_ORDERKEY P PARTIESY - 1. PARTIESY TABLE ACCESS PARTS

様子を、図 2に示す。横幅は火行時間、縦幅は LBA(数 説み込んだ LINEITEM のインデックス内にデータが Q8 の ORACLE8 による 次際のディスクアクセスの プロットした点と点を結ぶ線が、ディスクのシークを **単プロックアドレス)を示している。この図に於いて、** 存在していた為、テーブルにはアクセスしていない。

表していると考えることが出来る。

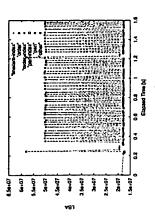


図 2: Q8 のディスクアクセス (最初の 1.6ms)

であるシーク時回の点から考えると、非常に効率の思 スする為、ディスクアクセス時間の支配的要素の一つ いアクセスが行われていると考えられる。また、ホス アクセスを行うみ、ホストがデータを受け取って、次 の命令を出すまでの頃、ディスクはストールした状態 ORDERS のインデックスとテーブルに交互にアグセ トは、過去にアクセスしたデータに基づき、ディスク 四2からも分かるように、Q8に関し、ディスクは、 となっている。(図3)

2.2 アクセスプランペースのプリフェッチ

セスプランがディスクに与えられた場合を考える。こ 欠に、アプリケーションフペケの対観とした、アク

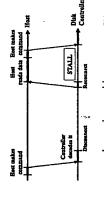


図 3: ディスクのストール

セスナペきなのか判断できるようになる。徐って、OR-さらに、データがディスクアレイのような複数台のディ 群をスケジューリングし、物理的に近い所からまとめ てアクセスする事により、シーク時間を短縮する事が に並行してアクセスできる為、さらなる実行時間の短 O場合、ディスクは、ORDERS のインデックスを設 出来る。また、このディスクアクセスは、ホストから んだ後、 ORDERS のテーブルのどのブロックにアク DERS のインデックスを連続的にアクセスし、そのイ ンデックスによりアクセスすべきデーブルのブロック の命令を待つ事なく行われる為、ホストとディスクは、 スクに分散して保存されている場合、個々のディスク 並行して処理作業を行う事ができる。(図4) 留が期待できる。

3 実ディスクアレイを用いた撥似実験

程度の性能向上が見込まれるのかを評価する為に、ディ ジューリングによるプリフェッチを行った場合、どの スクコントローラによるプリフェッチを擬似的に実現 するシステムを構築し、評価実験を行う。本来、ディ スクコントローラを高機能化し、プリフェッチ機構は ディスク上に実装されるべきものであるが、その得集 ディスクにおいてアクセスプラン情報を用いたスケ

į	Coche M	→ Ditt Cestrolier	1	l
Bert Best makes made dess, commune]	Prefetching	Transfer (In:
_		- Aguage	Jeresdos II	Bank rimes Remotes in Interest
Beef makes encounted	Kamistigs of	Oues Plan	•	

ORDERDATE BETWEEN

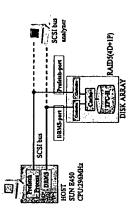
INDEX SCAN ORDER

図 4: アクセスプランを用いたディスクアクセス

は困難であることから、ホスト上にプリフェッチ機構 を擬似的に実行するプリフェッチプロセスを作成し、 代行させる。

3.1 実験環境

で、ディスクアレイのローデバイス上に RAID5(4D+1P) データサイズはScale Factor = 2、すなわち約2G(Byte) prefetch ポートと名付ける。 ORACLE 用の TPC-H の し、個別にコマンドを発行する事ができる。また、ディ 実験環境を図 5に示す。 SUN Ultra Enterprize450(4 DBMS において問合せを実行した場合、ディスクアレ イには1方のポートからアクセスが行われる。ここで、 × Ultra SPARC-II 296MHz) をホストとし、ディス クアレイ (Hitachi DF400) を接続する。 ディスクアレ イにはアクセス用のポートが2つあり、それぞれがホ DBMS がアクセスするポートを DBMS ポートと名付 で保存されている。ホストは、この2つのボートに対 スクのキャッシュは2 つのポートに於いて共有されて いる。実験に使用した TPC-H の各テーブルの構成を ストに Wide SCSI インタフェースでつながれている。 け、プリフェッチプロセスがアクセスするポートを 表 11に示す。



図で、シャリフーション戦机

また、ディスクアレイのそれぞれのボートへのアク セス状況は、SCSI パスアナライザを通し、PC によ

PARTS
1

表 1: 実験用 TPC-H データセット

జ	56	2711	6562	41003	68164	4913	123409	
	PARTS.INDEN	PARTS	LINEITEN INDEX	ORDERS INDEX	ORDERS	ОТНЕЯ	TOTAL	

表 2: 各テーブルへのアクセス数 (TPCH Q3)

り探取することが出来る。

ディスクアクセストレース 3.2

TPC-H において、ORACLE システムが Q8 を実行 **ずを用い採取した。 Q8 の実行時間は、1180 わであっ** た。また、Q8に関して行われた各テーブルへのアク した場合のアクセストレースを、 SCSI パスアナライ セス数を数2に示す。ORACLEシステムの最小アク セスブロック単位は 2k(Byte) であり、ほとんどのア クセスが最小プロック単位で行われていた。

ブリフェッチシステム

この節では、ディスクコントローラによるブリフェッ チを板似的に実現するプリフェッチシステムについて、 説明を行う。

ORACLE8 は DBMS ポートを通し、ディスクアク セスを行う。この時、修正されたデバイスドライバに より DBMS ポートからの全てのディスクアクセス信 報が、アクセスが発生すると同時にプリフェッチプロ セスに知らされる。プリフェッチプロセスは、このア クセスを監視し、 ORACLE が特定のプロックヘアク セスを行うと、トレース情報を基に、 prefetch ポート を通して、特来、ORACLE が参照すると考えられる

データのブリフェッチを実行する。ディスクアレイ上では、キャッシュが2つのボート間で共有されている為、ディスクコントローラがブリフェッチを行った場合と同じ状況が複似的に作成される。ブリフェッチンステムを因のに示す。

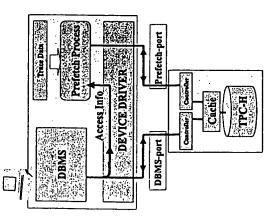


図 6: プリフェッチシステム

3.4 プリフェッチアルゴリズム

説明する。アクセスプランに於いて、NESTED LOOPS よって、一度に行うプリフェッチの故 (プリフェッチの JOIN が行われている部分の最内田に沿目した時、そ この笛では、プリフェッチのアルゴリズムについて の以公因がインデックスを用いたテーブルアクセスで の短縮の効果が期待できる。アクセスプランが多重の NESTED LOOPS JOIN になっている場合、効率的 なスケジューリングを行うという観点からは、より外 るまでディスクキャッシュから溢れない様に利用する 必要があり、ディスクキャッシュ智益の制限を受ける。 ある場合、その最内因の部分に於いて、アクセススケ ジューリングを用いたプリフェッチによるシーク時間 **囚包のループによるプリフェッチが凶ましいが、一奴** の先説みで得たデータを ORACLE システムが参照す 除さ) は、統計信頼などにより必要なディスクキャッ シュが甘を予朗し、決定される。

Q8を例に取り、説明することにする。図1のアクセ

実行時間(s)	1180	423	338	233
	プリフェッチ無し	r—> X∃QNI:W∃LINIT	FART ベース	PART.INDEX ベース

数 3: Q8 における実行時間の比較 (同時先説み IO 発作等 ao

スプランを考えた場合、Q8 は2 重の NESTED LOOPS
JOIN を構成している。最内周は、ORDERS.INDEX
を用いた ORDERS. へのブラセスであり、LINETEM.INDEX、
PARTS. PARTS.INDEX の条件館の基で、プラセス
される。Q8 の場合の長も続い先端みは ORACLE が
LINEITEM.INDEX はアクセスを行ったと同時に、そ
OLINEITEM.INDEX から得られる ORDERS.INDEX
と ORDERS のグリフェッチを行うことである。Q8 の
場合のブリフェッチのループの後補は、

- ORDERS.INDEX,ORDERS のプリフェッチ (LINEITEM.INDEX ベース)
- LINEITEM.INDEX,ORDERS.INDEX,ORDERS 07 17 x y ≠ (PARTS ~~x)
- PARTS, LINEITEM.INDEX, ORDERS. INDEX, ORDERS of J > 2 ≠ (PARTS.INDEX ~~~)
- 4. PARTS.INDEX,PARTS,LINEITEM.INDEX, ORDERS.INDEX,ORDERS 07 9 7 ± 27

となる。ここで、春やが小さい程、浅いブリフェッチを示し、春をがし、春らが大きい程、深いブリフェッチを示している。今回の実験では、Q8に関して、LINEITEM.INDEXペース(毎年1)、PARTS.INDEXペース(毎年3)、PARTS.INDEXペース(毎年3)のループに関して実験を行った。

4 実験結果

4.1 プリフェヴチの深さと実行時間

同時先路本10 発行数が30 の場合の LINEITEM.INDEX ペース、PARTS.INDEX ペースのプリフェッチを行った場合の実行時間を表 3に示す。数から明らかな様に、プリフェッチを行わない場合と比較して、プリフェッチを行った場合の実行時間が大きく短縮されていることが判る。また、プリフェッチを深くした場合について比較すると、プリフェッチの深さを深くした場合の方がより大幅な性能の向上が示されている。最も深いプリフェッチでは、プリフェッチ無しの場合の5 毎以上の性能向上が示されている。以下、アクセストレースを見ながら、より詳細な解析を行う。

4.2 プリフェッチの深さ

この節では、プリフェッチの様さの影響について考える。 同時先記み10 発行数が30 の場合の LINEITEM. INDEX ペース、 PARTS ペース、 PARTS INDEX ペースの DBMS ポート、 prefetch ポートからの 1/0 発行 トレースを図 7、 図 8(LINEITEM.INDEX ペース)、図 9、図 10(PARTS ペース)、図 11、図 12(PARTS ペース)、区 11、図 12(PARTS ペース)、に示す。 機械は時間を示し、 縦軸は LBAを示している。

5。PARTS.INDEX ベースの場合には、PARTS.INDEX 一度も発生しない。この様に先記みが深い程、プリフェッ これらの殴によって、 ORACLE がベースとなるブロッ LINEITEM.INDEX ヘアクセスを行うと阿時に、プリ LINEITEM.INDEX にアクセスを行うまでストールす クへのアクセスを行ったと同時に、prefetch ポートよ りプリフェッチが開始されていることが、魔器できる。 フェッチプロセスはその LINEITEM.INDEX から得 チを行う。よって、そのプロック群がプリフェッチさ ストールを示している。次に PARTS ベースの場合に ついて考える。PARTS ベースのブリフェッチを行っ た場合、今度は、 PARTS 年にプリフェッチ機能のス トールが起こる (図 10) が、LINEITEM.INDEX ペー 毎にストールするが、そのストールは PARTS ベース の時よりもさらに少なく、図 12の枠に入る時間内には られる ORDERS.INDEX と ORDERS のプリフェッ れた後、プリフェッチプロセスは、ORACLEが次の 5。図8に於いて、梅穣の部分がプリフェッチ機能の スの場合 (図8) と比べると、はるかに少ないことが判 **子磁能がストールする場合が少なくなり、性能が向上** LINEITEM.INDEX ペースの場合、ORACLE が

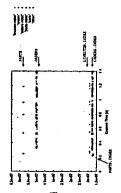


図 7: DBMS ポートのアクセス (LINEITEM.INDEX ベース)

4.3 キャッシュヒット状況

プリフェッチの祭さごとのアクセスのヒット状況に ついて検証する。



図 8. prefetch ポートのアクセス (LINEITEM.INDE.X ペース)



図 9: DBMS ポートのアクセス (PARTS ベース)

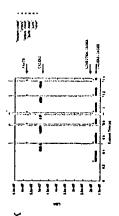


図 10: prefetch ポートのアクセス (PARTS ベース)



図 11: DBMS ポートのアクセス (PARTS.INDEX ベース)



図 12: prefetch ポートのアクセス (PARTS.INDEX (ドーツ

スのブリフェッチのヒット状況、ミス状況を、それぞ 歯磨やポした PARTS ペース、PARTS.INDEX ペー 化, 图13, 图14, 图15, 图1615示す。

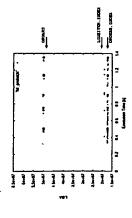
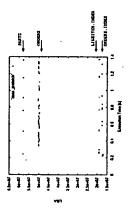


図 13: ヒット状況 (PARTS ベース)



対 14: ミス状況 (PARTS ペース)

図13、図15において、プリフェッチがヒットするこ 最初の約0.6 秒間は、DBMS ポートからは、PARTS とにより1/0時間が短縮され処理性能が向上している テーブルと LINEITEM.INDEX のプリフェッチが行 われているみ、プリフェッチの効果が現れていないが、 **ずがプロット点の集中により確認できる。図 16では、** その後、DBMS ポートのプリフェッチが ORDERS.

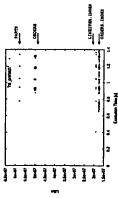


図 15: ヒット状況 (PARTS.INDEX ベース)

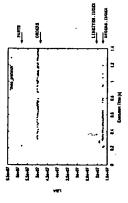


図 16: ミス状化 (PARTS.INDEX ペース)

INDEX と ORDERS に移る為、プリフェッチのヒッ ト串が上昇し、処理時間が短縮されている。 **数4に、それぞれの先説みの磔さに関するブリフェッ** チヒット母を示す。表から明らかな様に、プリフェッ チの欲さを深くする事により、ヒット母が向上してい ることが赴る。

4.4 同時先読みIO 発行数の影響

一度にディスクがスケジューリングを行う事の出来 る同時先試み10.発行数が、処理時間にどの程度、影 響を及ぼすのかを調べる為に、同時先説み10発行数 に対する処理時間をプロットした。その図を図 17に示 r。 機軸が同時先読み 10 発行数、縦軸が実行時間で

ヒット事(%)	0	۲ 77.4	۶2.8	۸ 85.8	
•	一先説み無し	TINEITEM.INDEX ベース	PARTS ベース	PARTS.INDEX ベース	

表 4: プリフェッチヒット単の比較 (同時先能み10 晃 **行数30**)

参考文献

35.

 D.Patterson et al. "Intelligent RAM(IRAM); the Industrial Setting, Applications, and Architectures". In Proceedings of the International Conference on Computer Design, 1997

PARTS INDEX 0436" ----

8 8 8 8 8 8

- [2] Erik Riedel, Garth Gibson, and Christors Faloutsos. "Active Storage For Large-Scale Data Mining and Multimedia". Proceedings of the 24th VLDB Conference, New York, USA, 1998
- [3] Anurag Acharya, Mustafa Uysal, and Joel Saltz. "Active Disk: Programming Model, Algorithms and Evaluation". In Proceedings of ASPLOS VIII, page81-91,Oct 1998
- M.Hellerstein. "A Case for Intelligent Disks(IDISKs)" [4] Kimberly Keeton, David A.Patterson, and Joseph SIGMOD Record, Volume 27, Number 3, August

Q8 に於いて、同時先點み IO 発行数が 1 から 10 ま での少ない危困では、発行数が10増えるごとに処理 時間の著しい削減がなされている。これは、ディスク アレイに於いて、ディスク5台にデータを分散して保 コマンドが発行されることによって実現される並列処 型による性能向上とスケジューリングの効果によるも の考えられる。しかし、同時先読み IO 発行数が 15 を 組えた辺りからは、PARTS.INDEX ベースでは若干 処理性能は勉和し、同時先説み IO 発行数の増加に伴

図 17: 同時先説み10 発行数に対する実行時間

Case for Redundant Arrays of Inexpensive Disks(RAID)" [5] D.A.Patterson, G.Gibson, and R.H.Katz. "A In Proceedings of ACMSIGMOD, pp.109-116, Jun, 1988

待している為、それぞれのディスクに対し、同時に read

[6] A.Acharya, M.Uysal, and J.Saltz. "Active disks" Technical Report, TPC98-06, University of California, Santa Barbara, March 1998

の性能向上が見られるが、その他の場合に於いては、

う処理時間の向上は見られない。

5 まとめと今後の課題

- "An Evaluation of Architectural Alternatives for ters, SMPs". Technical Report TRCS98-27, Uni- Mustafa Uysal, Anurag Acharya, and Joel Saltz. Rapidly Growing Datasets: Active Disks, Clusversity of California, Santa Barbara, Oct 1998
- [8] Cheetah Specification. http://www.seagate.com/
 - [9] Anurag Acharya, Mustafa Uysal, and Joel Saltz. "Structure and Performance of Decision Support Algorithms on Active Disks". Technical Report, TRCS98-28, University of California, Santa Barbara, Oct 1998

性節向上が得られるのかを示す為に、擬似的なプリフェッ

アクセスプランによるプリフェッチを行うことにより、 ヒット串が向上し、性能が大幅に改善される事が示さ

チシステムを構築し、評価実験を行った。その結果、

その他の DBMS においても同じような結果が終られ

れた。今回は、ORACLE8 に関して実験を行ったが、

るのか、また、その他の間合せに於いても同様な効果 が得られるのかなどの検討が必要と考えられる。今後 **は、これらの事を踏まえ、さらなる実装実験を行って**

プリフェッチ機構を持つディスクに於いてどの程度の

アプリケーションフベルの苔類として、アクセスプ ランが与えられた場合、そのアクセスプランを用いた

- David A.Patterson. "ISOTRE:Introspective Stor-Operating Systems (HotOS-VII), Rio Rico, Arizona [10] Aaron Brown, David Oppenheimer, Kimberly age for Data-Intensive Network Services". Proceedings of the 7th Workshop on Hot Topics in Keeton, Randi Thomas, John Kubiatowicz, and March 1999.
- 11] http://www.tpc.org/
- セスブランを用いたプリフェッチ磁降の一考額"、DEWS'99 [12] 稲見殿, 寛連川優. "高機能ディスクにおけるアク

本研究に御協力賜わりました、日立製作所の大枝氏、

松並氏に感謝致します。